A multi-task deep learning framework coupling semantic segmentation and image reconstruction for very high resolution imagery

Maria Papadomanolaki, Konstantinos Karantzalos, Maria Vakalopoulou

A MULTI-TASK DEEP LEARNING FRAMEWORK COUPLING SEMANTIC SEGMENTATION AND IMAGE RECONSTRUCTION FOR VERY HIGH RESOLUTION IMAGERY

Maria Papadomanolaki[1,2], Konstantinos Karantzalos[1], Maria Vakalopoulou[2]

[1] Remote Sensing Laboratory, National Technical University of Athens, Greece
[2] CVN, CentraleSupélec, Université Paris-Saclay and INRIA Saclay, France
mar.papadomanolaki@gmail.com, karank@central.ntua.gr, maria.vakalopoulou@centralesupelec.fr

ABSTRACT

Semantic segmentation, especially for very high-resolution satellite data, is one of the pillar problems in the remote sensing community. Lately, deep learning techniques are the ones that set the state-of-the-art for a number of benchmark datasets, however, there are still a lot of challenges that need to be addressed, especially in the case of limited annotations. To this end, in this paper, we propose a novel framework based on deep neural networks that is able to address concurrently semantic segmentation and image reconstruction in an end to end training. Under the proposed formulation, the image reconstruction acts as a regularization, constraining efficiently the solution in the entire image domain. This self-supervised component helps significantly the generalization of the network for the semantic segmentation, especially in cases of a low number of annotations. Experimental results and the performed quantitative evaluation on the publicly available ISPRS (WGIII/4) dataset indicate the great potential of the developed approach.

***Index Terms*—** Deep learning, Fully-convolutional networks, Feature representations, Autoencoders, Limited annotations

1. INTRODUCTION

Semantic segmentation is a very important field in several computer vision problems and among the most actively researched topics by the remote sensing community. During the last years, advances in deep learning have resulted in powerful models capable of detecting successfully several earth observation semantic categories on a variety of spectral and spatial resolution datasets. Various classification approaches based on deep learning have been proposed in the recent literature [1] in an effort to achieve better accuracies in a variety of applications and create robust and efficient detection systems.

Until today, fully convolutional networks, initially presented in [2] deliver the state-of-the-art results and produce the best accuracy rates in several semantic segmentation benchmark challenges [3, 4, 5]. Some of the most widely employed models for pixelwise semantic segmentation include fully convolutional architectures such as SegNet and U-Net together with a lot of variations. SegNet [6] is based on the encoder-decoder idea where the input is downsampled to a very low resolution and then upsampled back to its original dimensions. Similarly, U-Net architecture [7] follows the same idea adding skip connections between the encoder and the decoder allowing in this way the model to keep track of the different spatial resolutions and combine them in order to create more fruitful feature representations. Other approaches include residual learning [8, 9] which is also largely employed since it contributes to the elimination of the vanishing gradient problem when dealing with very deep architectures.

In this paper, we adopt a multi task deep learning based approach where semantic segmentation and image reconstruction processes are optimized simultaneously. Our assumption is that by solving jointly the two problems, using a common architecture with shared layers, we create more meaningful representations by keeping the image's properties leading in this way to higher accuracy especially when the annotated data are not sufficient. It should be mentioned here that the reconstruction task does not need additional annotations as it depends only on the raw image, however it can affect the observed parameters. A similar idea has been also tested on medical imaging [10] for the accurate detection of different brain tumor classes, using multisource medical volumes.

2. METHODOLOGY

Recent works report that multi-task learning can lead to high accuracy for the employed tasks, however, most of the times the needed annotations for the final optimization become n times more, where n indicates the number of simultaneously optimized tasks. In this paper, we investigate how a self-supervised task as image reconstruction can affect the semantic segmentation, enforcing better feature representations through the common layers. For our experiments we

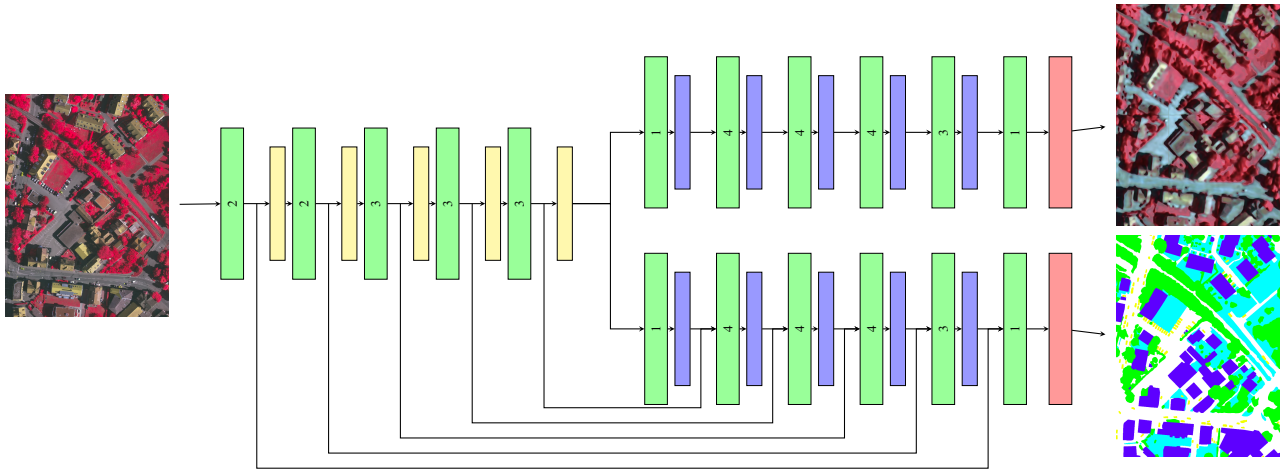

Fig. 1. The proposed U-REC architecture. Green layers: convolutional blocks with successive Conv,BN,ReLU operations. Yellow layers: max pooling, Blue layers: upsampling, Red layers: 1x1 convolution operations. The number inside each green layer indicates how many times such a convolution block is used. The final classification heatmap of the segmentation-related decoder-branch has a shape of *nClasses*x*W*x*H*, while the final reconstructed image has a shape of *nChannels*x*W*x*H*.

employed a fully-convolutional U-Net-like deep architecture (named U-REC) which has been proven to work well on semantic segmentation for very high resolution datasets. It should be mentioned here however that our framework can also be used with other types of architectures. The encoder part of the model consists of repetitive convolutional blocks which apply convolution, batch-normalization and rectified linear unit (ReLU) activation. Five max-pooling layers are used in total, bringing the input volume down to a very low resolution. After the encoding part the model is split into two decoding parts; one responsible for the segmentation and the other for the reconstruction task. The segmentation decoder branch follows the usual pattern of a U-Net architecture including convolutional blocks, upsampling procedures and concatenation operations between corresponding encoder-decoder parts. As far as reconstruction is concerned, the respective decoder branch consists of a similar layer succession, although this time skip operations are eliminated. This is essential in order to ensure that the model does not receive too much information about the original image through the skip connections, creating in this way a more constructive learning process. In Figure 1 one can observe the overall configuration of the architecture.

Under this framework, we have two different label types that need to be optimized: semantic segmentation labels l_{seg} with values $i = 1, 2, \ldots, K$, where K is the number of classes, and image reconstruction labels that correspond to the actual image spectral values. Each of these labels are optimised with a specific loss function. We chose the L1 norm to be the loss for the image reconstruction, minimizing the absolute difference of the true values of the image from the predicted ones. This way, L_1 loss is defined as,

$$L_1 = \sum_{i=1}^{n} |x_n - y_n| \tag{1}$$

where x_n indicates the true spectral values of the images, y_n represents the model's estimated output and n each of the pixels of the image.

Similarly, as we deal with a classification problem with more than two classes, we employed the multiclass cross entropy for the optimization of the semantic segmentation task. This way the L_2 loss is defined as,

$$L_2 = -\sum_{l=1}^{K} y_{s,l_{seg}} log(p_{s,l_{seg}}) \tag{2}$$

where $y_{s,l_{seg}}$ is a binary indicator that shows if class l is the correct answer for observation s and $p_{s,l_{seg}}$ holds the probability that observation s belongs to class l for K number of semantic classes. The final optimized loss can be summarized as follows,

$$L = w_1 \cdot L_1 + (1 - w_1) \cdot L_2 \tag{3}$$

where w_1 is manually defined.

2.1. Dataset and Implementation Details

All the experiments were conducted on the publicly available ISPRS (WGII/4) benchmark dataset depicting the city of Vaihingen. This dataset consists of 33 very high resolution images of average size 2494x2064 that have 3 available channels (Infrared, Red, Green) and a ground sample distance of 9cm. Six different classes are included, namely *Impervious Surfaces*, *Buildings*, *Low Vegetation*, *Trees*, *Cars* and *Clutter* which represents everything else that is not included in the other five classes. 14 out of the 33 images (areas 11, 13, 1,

21, 23, 26, 28, 30, 32, 34, 37, 3, 5 and 7) were used for training, 2 for validation (areas 15 and 17) and the rest for testing.

Patches of size 256x256 were extracted from the Vaihingen images using a step of 64 along both rows and columns forming in this way overlapping small regions. Approximately 13800 training patches were feedforwarded to the U-REC architecture during training and 120 validation ones were used for evaluation. Regarding hyperparameters, we chose the Adam optimizer while the batchsize and learning rate were equal to 14 and $1e^{-4}$ respectively. Moreover, after a grid search, the more appropriate value for the w_1 was defined to 0.1. Each epoch lasted about 10 minutes on a single NVIDIA GeForce GTX TITAN with 12 GB of GPU memory. All investigation trials were performed using the PyTorch library [11].

3. RESULTS AND DISCUSSION

The trained U-REC was evaluated on the 17 testing images of the ISPRS dataset. Results were compared with the plain U-Net architecture both on quantitative and qualitative terms. Beginning with the accuracy metrics, Tables 1 and 2 include the resulting confusion matrix of each method as well as precision, recall and F1 rates. As we can observe, similar results have been obtained in each case, with the U-REC architecture achieving higher F1 rates for all semantic categories except *Impervious_Surfaces*. It should be noted here that U-REC ameliorated greatly the *Clutter* category whose percentage in the training dataset is only 0.7%. This indicates the power of such an approach when dealing with limited annotated data. Overall Accuracy was also better since 89.02% and 88.60% resulted from U-REC and U-Net respectively.

In Figure 2 the qualitative evaluation of the employed method is provided. In the first row we can observe that *Trees* are better detected in the case of U-REC whereas the simple

Predicted \ Reference	imp_surf	building	low_veg	tree	car	clutter
imp_surf	**0.924**	0.035	0.027	0.010	0.004	0.000
building	0.038	**0.941**	0.015	0.005	0.001	0.000
low_veg	0.049	0.017	**0.773**	0.161	0.000	0.000
tree	0.011	0.002	0.075	**0.911**	0.000	0.000
car	0.146	0.073	0.006	0.002	**0.772**	0.001
clutter	0.281	0.441	0.005	0.003	0.088	**0.181**
Precision/Correctness	0.905	0.933	0.851	0.843	0.782	0.939
Recall/Completeness	0.924	0.941	0.773	0.911	0.772	0.181
F1	**0.915**	**0.937**	**0.810**	**0.876**	**0.777**	**0.303**

Table 1. Confusion matrix of the plain U-Net architecture.

Predicted \ Reference	imp_surf	building	low_veg	tree	car	clutter
imp_surf	**0.927**	0.027	0.033	0.011	0.002	0.000
building	0.042	**0.936**	0.015	0.006	0.000	0.000
low_veg	0.043	0.014	**0.773**	0.169	0.000	0.000
tree	0.008	0.002	0.058	**0.932**	0.000	0.000
car	0.204	0.062	0.007	0.005	**0.716**	0.006
clutter	0.369	0.315	0.014	0.010	0.035	**0.257**
Precision/Correctness	0.904	0.946	0.861	0.838	0.893	0.889
Recall/Completeness	0.927	0.936	0.773	0.932	0.716	0.257
F1	**0.915**	**0.941**	**0.815**	**0.882**	**0.795**	**0.399**

Table 2. Confusion matrix of the U-REC architecture.

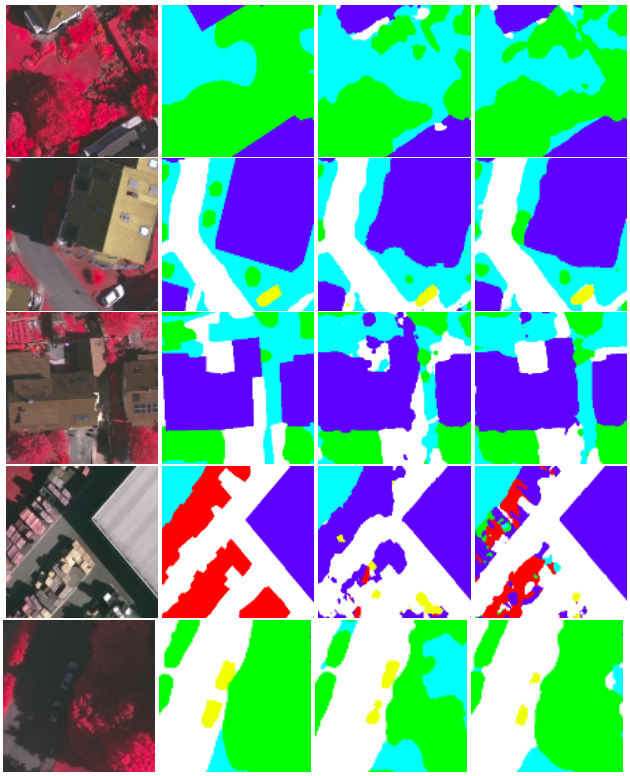

Fig. 2. Results from zoomed areas of testing images. Left to right: original image, ground truth, U-Net, U-REC. (White: *Impervious_Surfaces*, Blue: *Buildings*, Light Blue: *Low_Vegetation*, Green: *Trees*, Yellow: *Cars*, Red: *Clutter*)

U-Net architecture confuses them with *Low_vegetation*. This is also the case in the second row where single trees covered by the building shadow have been successfully spotted compared with U-Net who fails to recognize them. Continuing to the third row, one can notice that the *Impervious_Surface* which is enclosed inside the *Building* has been identified more appropriately in the case of U-REC. The fourth row also shows the employed architecture's superiority on the *Clutter* category. U-Net has failed almost completely to distinguish it from other semantic categories in contrast with U-REC which has detected correctly a large amount of *Clutter* pixels. It should be noted here that even though the U-REC learned to take advantage of feature representations more constructively in certain cases, its performance is inferior to U-Net for the *Cars* semantic class. This is evident from the last row of Figure 2 where U-REC was unable to identify cars existing under the tree shadow as opposed to U-Net.

The U-REC architecture's behaviour is inextricably related to the reconstruction learning process that has taken place during training. This is because the two decoder branches have many layers in common, thus the weights of the model are formulated based on both optimization procedures, namely semantic segmentation and reconstruction. It is therefore reasonable for the model to behave better in categories that the reconstruction was more successful. As we

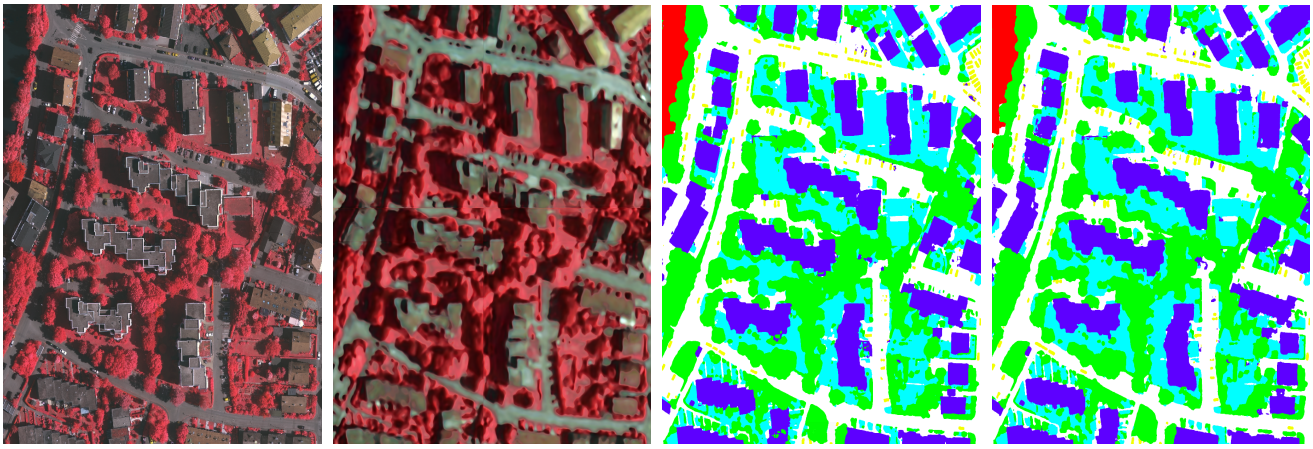

Fig. 3. From left to right: original testing image, reconstruction, U-REC, U-Net. (White: *Impervious_Surfaces*, Blue: *Buildings*, Light Blue: *Low_Vegetation*, Green: *Trees*, Yellow: *Cars*, Red: *Clutter*)

can observe in Figure 3, *Roads* and *Trees* are reconstructed more properly comparing to buildings and cars which have not been formulated with much detail. This can also indicate the the L1 norm is not the best loss for the accurate reconstruction of satellite data.

4. CONCLUSION

In this paper we investigated the coupling of semantic segmentation with image reconstruction using a multi-task scheme. Our assumption is that through the joint optimization of the two problems, the employed network will create better and more rich representations. Results on a publicly available dataset indicate that in general the reported accuracy of the different semantic labels can be boosted even more than state-of-the-art deep learning architectures. In the future, we aim to perform more experiments with different models for the pixelwise semantic segmentation of very high resolution datasets. Moreover, we are planning to investigate different task specific losses towards this direction.

5. REFERENCES

[1] X. X. Zhu, D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, 2017.

[2] Evan Shelhamer, Jonathan Long, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[3] Yansong Liu, Sankaranarayanan Piramanayagam, Sildomar T. Monteiro, and Eli Saber, "Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order crfs," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1561–1570, 2017.

[4] B. Huang, K. Lu, N. Audebert, A. Khalel, Y. Tarabalka, J. Malof, A. Boulch, B. Le Saux, L. Collins, K. Bradbury, S. Lefvre, and M. El-Saban, "Large-scale semantic classification: Outcome of the first year of inria aerial image labeling benchmark," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018.

[5] Maria Papadomanolaki, Maria Vakalopoulou, and Konstantinos Karantzalos, "A novel object-based deep learning framework for semantic segmentation of very high-resolution remote sensing data: Comparison with convolutional and fully convolutional networks," *Remote Sensing*, vol. 11, pp. 684, 03 2019.

[6] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 2481–2495, 2017.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, Eds., Cham, 2015, pp. 234–241, Springer International Publishing.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[9] Lichao Mou and Xiao Zhu, "Vehicle instance segmentation from aerial image and video using a multi-task learning residual fully convolutional network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, 07 2018.

[10] Andriy Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018*.

[11] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.